UDC 551.509.313:517.93

COMPUTATIONAL STABILITY AND TIME TRUNCATION OF COUPLED NONLINEAR EQUATIONS WITH EXACT SOLUTIONS

F. BAER and T. J. SIMONS [1]

Department of Atmospheric Science, Colorado State University, Fort Collins, Colo.

ABSTRACT

A general numerical integration formula is presented that generates many of the commonly used one-dimensional finite-difference schemes. A number of these schemes are tested on a simple wave equation; three implicit and three explicit are chosen for further analysis with a nonlinear set of equations with known solutions. A seventh method of the implicit type not requiring iteration is also tested. A transformation is developed that allows the removal of linear terms from the nonlinear equations, thereby avoiding truncation of the linear terms. The results of the analysis show that energy components may have large errors when the total energy shows essentially none, and phase errors may be quite serious without indication from linear analysis. By treating the uncoupled linear terms exactly (no truncation), significant improvement in the numerical solutions ensues. The multilevel implicit schemes give superior results and are to be recommended if computing time is not a criterion. Great care must be taken in interpreting the linear stability criterion. The critical truncation increment should be considerably reduced to avoid significant truncation errors, especially for long time integrations.

1. INTRODUCTION

The problems of computational stability and truncation errors are by no means recent in origin. Indeed, few physical problems are so simple as to yield mathematical representations that lend themselves to analytic solutions. More often than not, the appropriate equations are nonlinear and must be solved numerically with little insight into the exact solutions. One is generally confronted with partial differential equations to further complicate clarification of the errors arising from numerical computation.

Despite these seemingly overwhelming obstacles, significant progress in studies of computational stability have been made as exemplified by the work of Richtmyer (1957). The traditional approach to such studies is to linearize the nonlinear equations and then compare the exact solutions of the linear system to the solution of the corresponding finite-difference equations. For different truncation procedures, the approximations may be evaluated in terms of the true solution. For initial value problems where the linearizing assumption may not be valid for all time, only the criterion of computational stability has utility. Moreover, since finite-difference operations must generally be applied in both space *and* time, highly involved relationships between the truncation intervals evolve.

With reference to problems concerning atmospheric flow, the feasibility of converting the appropriate nonlinear partial differential equations to a finite set of ordinary nonlinear first-order differential equations in time (termed "spectral" equations) has been established. Such equations are generated by assuming the space dependence to be given by a series of known polynomials and solving for the time dependent coefficients through integration over the entire space domain. The technique seems to have been applied first by Silberman (1954) and discussed in detail by Platzman (1960). On the assumption that the series truncation does not create serious errors (a question not yet investigated in detail) or that the finite set of equations is an exact representation of the physical system, one is left with the considerably simpler problem of determining time truncation alone.

The investigation of ordinary differential equations by numerical methods has also not been neglected; see, for example, Henrici (1962), Hildebrand (1956), or Milne (1949). However, if wave-type solutions exist, error estimates of linear equations based on Taylor expansions may be cast into doubt, and investigations of the type carried out by Kurihara (1965) are necessary. Little is known about truncation errors of initial value problems involving nonlinear equations. Fortunately, there exist some nonlinear systems of spectral equations that have analytic solutions. Such systems were first used to describe atmospheric flow by Lorenz (1960). Clearly, a comparison between the finite-difference solution of the equations of such a system when compared to the analytic solution will give information on truncation errors as a function of time. Studies with various time-differencing schemes have been made on this basis by Lilly (1965) and Young (1968).

A number of finite-differencing schemes have been utilized for integrating ordinary differential equations, and many are a composite of ingenious techniques that have occurred to various scientists and have been proven useful. For testing the utility of such schemes, however, it seems worthwhile to generate them in some systematic fashion, thereby establishing a hierarchy of schemes with (we hope) increasing accuracy. One such systematic

[1] Present affiliation, Canada Center for Inland Waters, Department of Energy, Mines, and Resources, Burlington, Ontario.

approach would be to assume that the function to be integrated can be represented by a polynomial that is exact at its known point values. The degree of accuracy of such a polynomial will then be established by the number of known points utilized. We shall show, moreover, that the most popular schemes can be represented by this approach.

The schemes were tested by application to a first-order linear wave equation to avoid the problem of being overwhelmed by an unmanageably large number of schemes. When a scheme was not able to give good results for this equation, we assumed it would not be satisfactory for a more complex system of equations. In this way, we were able to reduce the number of schemes to a manageable size. It should be noted that, if more points on the time axis are used to develop the interpolation polynomial than there are orders of derivatives in the differential equations, spurious solutions will result—frequently denoted as "parasitic" solutions—that must be handled with great care so as not to obscure the true physical solution.

The remaining schemes (those which gave satisfactory results with the wave equation) were then tested on a low-order spectral system of the type used by Lilly and Young. The system used here however (Baer 1970) has the added flexibility of involving both linear and nonlinear terms in the first-order system of equations; it furthermore allows for time-dependent phase changes that were constrained in previous experiments. Since linear contributions to differential equations may be determined without truncation, their influence has been investigated. Of the techniques that proved most accurate, multistep methods were included, despite the presence of parasitic solutions. Previous calculations suggest that integral constraints of the system (say, energy or vorticity) were adequate indicators of truncation error when observed during calculation. This conclusion does not seem to be borne out. We shall see that slight phase errors will create amplitude errors in the individual dependent variables that have a tendency to cancel when the integral properties are evaluated. Thus, although the integral constraints will yield a good indication of computational stability (which is also available from linear theory), truncation errors can only be investigated from the detailed behavior of all the dependent variables in the system.

2. TRUNCATION SCHEMES

As we have indicated, the spectral equations applicable to the atmosphere may be represented quite generally by a nonlinear set of first-order differential equations in time for which analytic solutions are not available unless the set is highly truncated. The dependent variables, which are the expansion coefficients of the space-dependent polynomials, may be represented by a vector $\boldsymbol{\Psi}$ such that

$$\boldsymbol{\Psi}=(\psi_i), \quad 1\leq i\leq N;$$

and the general set of equations may be written as

$$\dot{\boldsymbol{\Psi}}=\mathbf{F}(\boldsymbol{\Psi}, t)=(f_i) \tag{1}$$

where $\mathbf{F}$ is a vector operator and the dot notation signifies time differentiation. Suppressing indices, we may also state that the scalar equation for any expansion coefficient will be

$$\dot{\psi}=f(\boldsymbol{\Psi}, t). \tag{1a}$$

Because exact solutions in time are not available for equation (1) based on the complicated nature of the functions f, we may expect to know $\boldsymbol{\Psi}$ only at discrete points on the time axis. For simplicity, let us assume $\boldsymbol{\Psi}$ (and therefore $\mathbf{F}$) known at equal time increments

$$t=t_0+j\Delta t$$

and $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ (2)

$$j=0, 1, 2 \ldots, \tau.$$

Over a given interval in time, we may establish by an interpolation formula a continuous function of time that corresponds to the known values at the discrete points given by equation (2). If we consider the continuous variable in time to be given as

$$t=t_0+(\tau+s)\Delta t, \quad -n\leq s\leq 1 \text{ and } n\leq\tau, \tag{3}$$

then by Newton's backward interpolation polynomial (see, for example, Hildebrand 1956 or Milne 1949)

$$f(t)=\sum_{k=0}^{n}(-1)^k\binom{-s}{k}\Delta^k f^\tau. \tag{4}$$

In equation (4), the quantity in the second parentheses is a binomial coefficient function of s, the superscript on f denotes the increment in time at which the function should be evaluated (from equation 2, the function *is* known at the time τ), and Δ^k represents the backward difference operator applied k times and has the value

$$\Delta f^\tau\equiv f^\tau-f^{\tau-1}$$
$$\Delta^k f^\tau=\sum_{j=0}^{k}(-1)^j\binom{k}{j}f^{\tau-j}. \tag{5}$$

Although we have specified that f is known at $\tau+1$ points, we need not utilize all these values in establishing our polynomial (4), and hence we choose merely the last n point values. We may determine how the interpolation polynomial depends on the discrete point values by substituting (5) into (4) and noting the following identities:

$$\sum_{k=0}^{n}\sum_{j=0}^{k}=\sum_{j=0}^{n}\sum_{k=j}^{n}$$

and

$$\binom{-s}{j}\binom{-s-j}{k-j}=\binom{-s}{k}\binom{k}{j}$$

The polynomial becomes

$$f(s)=\sum_{j=0}^{n}\alpha_{Ej}(s)f^{\tau-j}$$
$$\alpha_{Ej}(s)\equiv\binom{-s}{j}\sum_{k=0}^{n-j}(-1)^k\binom{-s-j}{k}. \tag{6}$$

If we wish to establish the value of f at the point $\tau+1$, it would be necessary to extrapolate from (6); therefore we have used the subscript notation E. We could, however, *assume* the function known at $\tau+1$ and write an equation similar to (6) that would then allow an interpolation to the point $\tau+1$ and would read

$$f(s)=\sum_{j=0}^{n}\alpha_{Ij}(s)f^{\tau+1-j}$$
$$\alpha_{Ij}(s)\equiv\binom{-s+1}{j}\sum_{k=0}^{n-j}(-1)^k\binom{-s-j+1}{k}. \tag{7}$$

To establish the value of ψ $(\tau+1)$, we may now substitute either (6) or (7) into (1a) and integrate. The integration may go over any subinterval of the interpolation polynomial, but clearly not from a time preceding the point $\tau-n$. Selecting the integer p $(p\leq n)$ at which point the function is known and integrating to $\tau+1$, we have

$$\psi^{\tau+1}=\psi^{\tau-p}+\Delta t\int_{-p}^{1}f(s)ds. \tag{8}$$

It is interesting to note that use of the extrapolating polynomial yields an explicit solution for $\psi^{\tau+1}$, whereas the application of the interpolating polynomial leads to an implicit solution because the unknown function $f^{\tau+1}$ still exists on the right-hand side of the equation. If we define the integrals over s,

$$\int_{-p}^{1}\alpha_{Ej}(s)ds\equiv\bar{\alpha}_{Ej}(p)$$

and

$$\int_{-p}^{1}\alpha_{Ij}(s)ds\equiv\alpha_{Ij}(p), \tag{9}$$

where the integrals may be evaluated by noting that the integrals are factorial polynomials in s that may be converted to polynomials in s by use of Sterling's numbers of the first kind (Milne 1949), we may express the general finite-difference extrapolation formulas as

explicit, E_{pn}

$$\psi^{\tau+1}=\psi^{\tau-p}+\Delta t\sum_{j=0}^{n}\bar{\alpha}_{Ej}f^{\tau-j}$$

and

implicit, I_{pn} (10)

$$\psi^{\tau+1}=\psi^{\tau-p}+\Delta t\sum_{j=0}^{n}\bar{\alpha}_{Ij}f^{\tau+1-j}.$$

We see from (10) that a wide variety of finite-difference integration schemes may be selected in a systematic fashion. As we increase p and n, we arrive not only at higher order schemes (more "steps") with the consequent expected increase in accuracy but also additional parasitic roots. Most of the standard numerical integration schemes fall into the classification given by (10). For example, the schemes E_{0n} and E_{1n} are generally associated with Adams-Bashforth and Nystrom, respectively (Henrici 1964); whereas the schemes I_{0n}, I_{1n} are referred to as Adams-Moulton and Milne-Simpson, respectively (Hildebrand 1956). The more involved predictor-corrector or multicorrector schemes would require a sequence of schemes described by (10).

A number of schemes whose properties will be investigated are listed in table 1. Certain omissions will be noted. The E_{00} scheme, which is termed the "Euler forward" is always unstable in terms of fictitious amplification and is consequently of no interest. Similarly, the "Euler backward" I_{00} gives fictitious damping and is therefore ignored. Schemes with $p=2$ have been shown (Hildebrand 1956) to yield results not appreciably superior to those for $p=1$; their discussion would thus be redundant. For the implicit schemes, the coefficients $\bar{\alpha}_{I3}$ (1), $\bar{\alpha}_{I5}$ (3) vanish, and consequently the lower order forms I_{12}, I_{34} that require as much calculation as I_{13}, I_{35} have been ignored.

The schemes described by (10) may be subject to Taylor's series expansion about the point τ; for a given truncation (p, n), there will be an error of order $(\Delta t)^{n+2}$ times the same order of time derivative of ψ listed in table 1. We shall see in the sequel that applying this technique to wave-type equations may lead to misleading error estimates.

3. LINEAR STABILITY PROPERTIES

If the schemes listed in table 1 do not show adequate stability properties when applied to a linear differential equation, we may anticipate their failure with regard to nonlinear differential equations. We shall therefore test them on the simple linear wave equation

$$\dot{\psi}=-i\rho\psi \tag{11}$$

that could be generated from (10) by linearization and neglecting coupling terms. Note that ψ is a complex variable, but let us assume ρ to be real. The true solution of equation (11) shows only one mode that moves about the unit circle in the complex plane with period $2\pi/\rho$ beginning at unity when $t=2m\pi/\rho$. If we now define coefficients

$$\alpha_j\equiv\begin{cases}\bar{\alpha}_{I,j+1}, & \bar{\alpha}_{I,j+1}=0 \text{ for } j=n\\ \bar{\alpha}_{Ej}, & \bar{\alpha}_{Ej}=0 \quad \text{for } j=-1,\end{cases}$$

we may write both the implicit and explicit finite-difference schemes (10) after substitution of (11) for the values of the derivatives at the known discrete points by the single relation

$$(1+i\alpha-1\rho\Delta t)\psi^{\tau+1}=\psi^{\tau-p}-i\rho\Delta t\sum_{j=0}^{n}\alpha_j\psi^{\tau-j}. \tag{12}$$

TABLE 1.—*Values of the coefficients $\overline{a}_{Ij}(p)$, $\overline{a}_{Ej}(p)$ for different integration schemes E_{pn}, I_{pn}, their names (if known), and the truncation error based on Taylor's series analysis*

Scheme (p, n)	Name	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	Truncation error
E_{01}		3/2	−1/2					$\lvert 5/12(\Delta t)^3\psi^{(3)}\rvert$
E_{02}	Adams-Bashforth	23/12	−4/3	5/12				$\lvert 3/8(\Delta t)^4\psi^{(4)}\rvert$
E_{03}		55/24	−59/24	37/24	−9/24			$\lvert 1/3(\Delta t)^5\psi^{(5)}\rvert$
E_{04}		1901/720	−2774/720	2616/720	−1274/720	251/720		$\lvert 1/5(\Delta t)^6\psi^{(6)}\rvert$
E_{11}	Leapfrog	2	0					$\lvert 1/3(\Delta t)^3\psi^{(3)}\rvert$
E_{12}		7/3	−2/3	1/3				$\lvert 1/3(\Delta t)^4\psi^{(4)}\rvert$
E_{33}	Milne predictor	8/3	−4/3	8/3	0			$\lvert 1/3(\Delta t)^5\psi^{(5)}\rvert$
I_{01}	Euler trapezoidal	1/2	1/2					$\lvert 1/12(\Delta t)^3\psi^{(3)}\rvert$
I_{02}		5/12	8/12	−1/12				$\lvert 1/24(\Delta t)^4\psi^{(4)}\rvert$
I_{03}	Moulton corrector	9/24	19/24	−5/24	1/24			$\lvert 1/36(\Delta t)^5\psi^{(5)}\rvert$
I_{04}		251/720	646/720	−244/720	106/720	−19/720		$\lvert 1/53(\Delta t)^6\psi^{(6)}\rvert$
I_{13}	Milne corrector	1/3	4/3	1/3	0			$\lvert 1/90(\Delta t)^5\psi^{(5)}\rvert$
I_{14}		29/90	124/90	24/90	4/90	−1/90		$\lvert 1/3(\Delta t)^6\psi^{(6)}\rvert$
I_{35}	Milne II corrector	14/45	64/45	24/45	64/45	14/45	0	$\lvert 1/120(\Delta t)^7\psi^{(7)}\rvert$

The solutions to equation (12) may be determined in a number of ways, but they must all satisfy the characteristic equation

$$(1+i\alpha-1\rho\Delta t)\lambda^{n+1}=\lambda^{n-p}-i\rho\Delta t\sum_{j=0}^{n}\alpha_j\lambda^{n-j} \tag{13}$$

where the roots of (13) represent the solutions of (12). Since we have specified $p\leq n$, there will be $n+1$ solutions to (12), only one of which corresponds to the real "physical" mode. The computational or parasitic modes (n of them) are distributed as follows at $\Delta t=0$; $n-p$ roots begin at the origin, and $p+1$ roots are distributed equally about the unit circle with the physical mode at $\lambda=1$. As Δt is increased from zero, the roots will change from their initial points.

If the first root, λ_0, represents the "computed physical mode," we may compare it with the true solution. So long as its amplitude remains near unity, there will be no spurious damping or amplification. However, its phase, say θ_0, must also remain near the true phase for accuracy; that is, we should observe that $-\theta_0/\rho\Delta t$ remains close to unity. The remaining n solutions are parasitic and enter only to disturb the physical solution. So long as their amplitudes remain less than unity (that is, within the unit circle), they will be damped. If they go outside the unit circle, they will cause amplification and may be classified as "unstable" solutions. If they remain on the unit circle, by suitable choice of initial conditions their effects can be made innocuous.

All the schemes listed in table 1 have been tested on equation (11). Their characteristic equations may be easily determined by substitution of the tabular coefficients together with the limits (p, n) into (13). The roots of these characteristic equations have been determined for various values of $\rho\Delta t$, and the amplitudes of all modes for each scheme have been plotted against $\rho\Delta t$ (abscissa) in figure 1. Pursuant to the previous discussion, wherever a mode exceeds unity on the ordinate, it will yield an unstable solution. Clearly, the best schemes will be those for which all roots remain stable for the largest value of $\rho\Delta t$.

We may be considerably more precise about the behavior of these schemes by investigating the computed physical mode in more detail—both its amplitude and phase. On figure 2, we have plotted for all schemes the amplitude of the computed physical mode (and amplitudes of parasitic modes when they are within the ordinate scale) on the upper graph and the ratio $-\theta_0/\rho\Delta t$ on the lower graph against $\rho\Delta t$ on the abscissa. Here, we may isolate the best schemes. Whereas from figure 1 we might have thought that scheme I_{01} was best because it is stable for all values of Δt, we see from figure 2 that this scheme (trapezoidal) has serious phase errors for reasonable values of $\rho\Delta t$.

We have selected, based on figure 2, three schemes in the explicit group and three in the implicit group for further study. Choosing $\rho\Delta t\leq 0.4$, we see that E_{03}, E_{11}, and E_{33} are the best; whereas for the implicit schemes, the obvious choices are I_{01}, I_{13}, and I_{35}. Scheme I_{01} was selected because of the strong stability property of its amplitude and also because of its general popularity, although its phase characteristics are less desirable.

An interesting sidelight to the selection of suitable finite-difference schemes is exemplified by figure 3. Suppose one would like a scheme no greater than two-step for which the coefficients could be varied such that the most favorable properties may be chosen. Let the scheme be represented in terms of the arbitrary coefficients (α, β)

$$\psi^{\tau+1}=\psi^{\tau-1}+\Delta t(\alpha\dot{\psi}^{\tau-1}+\beta\dot{\psi}^{\tau}+\alpha\dot{\psi}^{\tau+1}) \tag{14}$$

$$2\alpha+\beta=2$$

and test it on the equation $\dot{\psi}=-i\nu\psi$. Figure 3 shows for various combinations of α, β that the phase properties in the stable range (where both the real physical and the parasitic roots have amplitude unity) are effectively bounded by the error curves for the leapfrog (L-E_{11}) on the one hand and the trapezoidal (T-I_{01}) on the other. The Milne scheme (M-I_{13}) is undoubtedly one of the best that satisfies the criteria of (14).

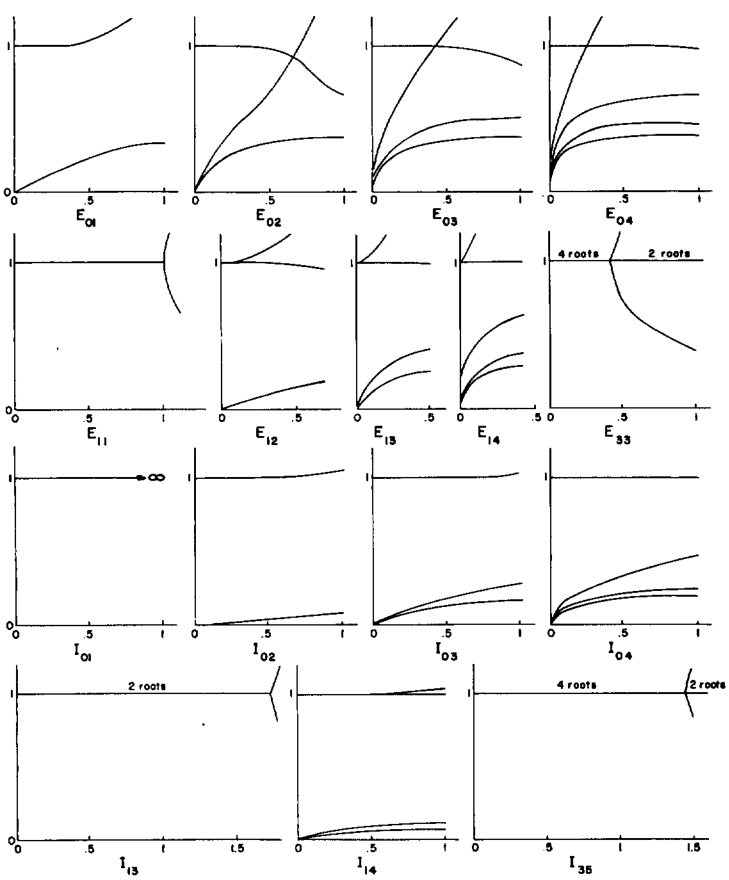

FIGURE 1.—Amplitudes of the roots of the simple linear wave equation $\dot\psi=i\rho\psi$ for various truncation schemes plotted against $\rho\Delta t$ on the abscissa.

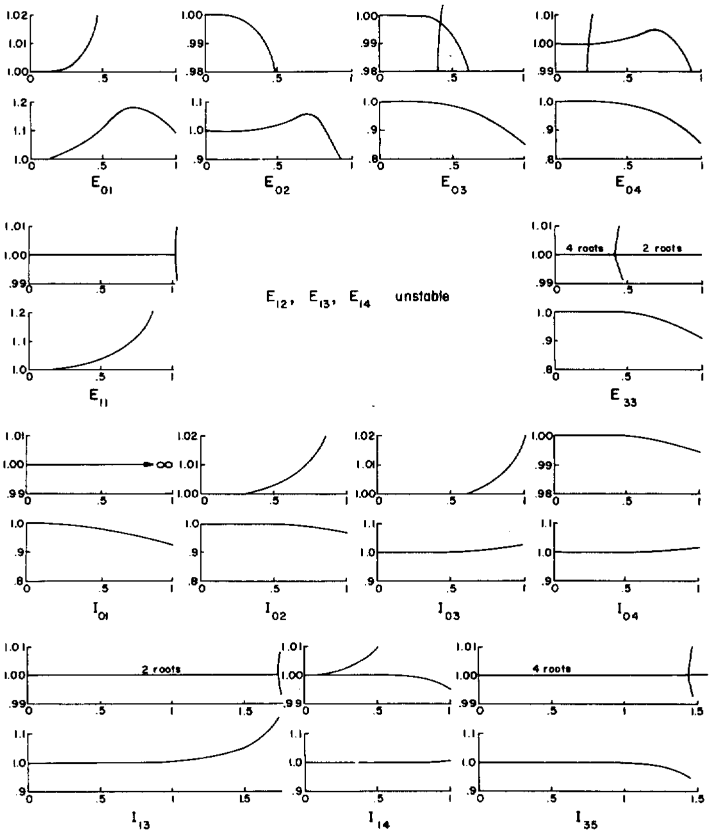

FIGURE 2.—Amplitude (upper) and phase (lower) of the physical root of the truncated linear wave equation plotted against $\rho\Delta t$. The phase is given proportional to $\rho\Delta t$.

4. MULTICOMPONENT SYSTEM

The coupled set of nonlinear first-order differential equations on which the six schemes that survived the linear analysis of the last section will be tested is part of the group of low-order spectral systems which were sytematically developed by Platzman (1962) for the barotropic vorticity equation, but which also have applicability for baroclinic problems. The system under consideration involves an arbitrary zonal flow interacting with a single planetary wave composed of two complex components (describing its latitudinal variability) in a rotating atmosphere with spherical geometry. Details of this system including the exact solutions (elliptic functions) have been presented by Baer (1970). If the zonal coefficients (real) are denoted as $\psi_\gamma(t)$ where $\gamma=2m+1$, $m\leq M$, and the complex wave coefficients are described by the terms $\psi_\alpha(t)$, $\psi_\beta(t)$, the differential equation for the zonal terms may be written

$$\dot\psi_\gamma=2a_\gamma \operatorname{Im}\psi_\alpha\psi_\beta^*.$$

The zonal coefficients can be solved in terms of one coefficient ψ_n by integration of the above equation. The time relationship thus developed between the zonal coefficients is unaltered if the integration is performed by numerical means, whereby we find that $\psi_\gamma=(a_\gamma/a_n)\psi_n+s_\gamma$. The system to be integrated therefore involves only three variables, ψ_n, ψ_α, and ψ_β and is

$$\begin{aligned}&\dot\psi_n=2a_n \operatorname{Im}\psi_\alpha\psi_\beta^*,\\ &\dot\psi_\alpha=-i\rho_\alpha\psi_\alpha+ih_{\alpha\beta}\psi_\beta+ig_{\alpha\alpha}\psi_n\psi_\alpha+ig_{\alpha\beta}\psi_n\psi_\beta,\\ \text{and}\quad &\dot\psi_\beta=-i\rho_\beta\psi_\beta+ih_{\beta\alpha}\psi_\alpha+ig_{\beta\beta}\psi_n\psi_\beta+ig_{\beta\alpha}\psi_n\psi_\alpha\end{aligned}\tag{15}$$

where $\rho_{\alpha,\beta}\equiv\nu_{\alpha,\beta}-h_{\alpha\alpha,\beta\beta}$. In matrix notation, we find

$$\begin{aligned}&\dot\psi_n=\tilde{\mathbf{\Psi}}^*\mathbf{H}\mathbf{\Psi}, && \psi_n \text{ real scalar},\\ &\dot{\mathbf{\Psi}}=(\mathbf{A}+\psi_n\mathbf{D})\mathbf{\Psi}, && \mathbf{\Psi}=\begin{pmatrix}\psi_\alpha\\ \psi_\beta\end{pmatrix},\\ &\mathbf{A}=\mathbf{A}_1+\mathbf{A}_2, && \psi_{\alpha,\beta}=\frac{1}{\sqrt{2}}B_{\alpha,\beta}(t)e^{i\theta_{\alpha,\beta}(t)},\\ &\mathbf{A}_1\equiv-i\begin{pmatrix}\rho_\alpha & 0\\ 0 & \rho_\beta\end{pmatrix}, && \mathbf{A}_2\equiv i\begin{pmatrix}0 & h_{\alpha\beta}\\ h_{\beta\alpha} & 0\end{pmatrix},\quad \mathbf{D}\equiv i\begin{pmatrix}g_{\alpha\alpha} & g_{\alpha\beta}\\ g_{\beta\alpha} & g_{\beta\beta}\end{pmatrix},\\ &\mathbf{H}\equiv ia_n\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}=-\tilde{\mathbf{H}},\end{aligned}\tag{16}$$

where the tilde denotes transposition. The physical significance of the constants that depend on spectrum

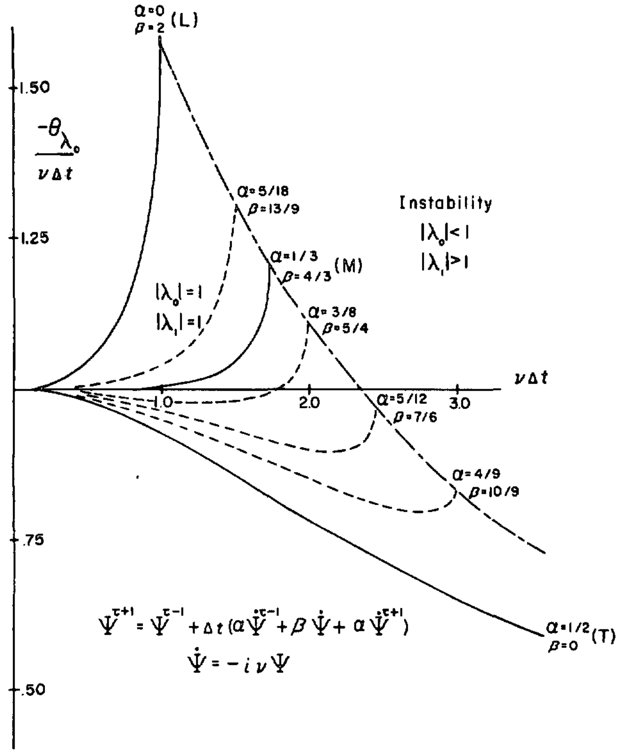

FIGURE 3.—Phase errors for a class of two-level truncation schemes in terms of the true phase including the leapfrog, trapezoidal, and Milne schemes.

truncation and initial conditions may be found in Baer (1970). Three different sets of conditions were used in this study, and the numerical values of the variables may be found in table 3 in section 5.

System (16) involves both linear and nonlinear terms; the uncoupled linear effects are denoted by the matrix $\mathbb{A}_1$, and the purely coupled linear terms by the matrix $\mathbb{A}_2$. The term $\psi_n\mathbb{D}\boldsymbol{\Psi}$ represents the nonlinear effect. The linear terms may be removed from truncation in a numerical integration by recognizing their exact influence. Let us therefore define the vector

$$\mathcal{X}\equiv e^{-\mathbb{G}t}\boldsymbol{\Psi} \tag{17}$$

such that the second differential equation of (16) becomes

$$\dot{\mathbb{X}}\equiv e^{-\mathbb{G}t}(\mathbb{A}-\mathbb{G}+\psi_n\mathbb{D})e^{\mathbb{G}t}\mathbb{X} \tag{18}$$

where $\mathbb{G}$ is a matrix that we shall allow to take on the three possible values

a. $\mathbb{G}=\mathbb{0}$, (TL)
b. $\mathbb{G}=\mathbb{A}_1$, (TC)
c. $\mathbb{G}=\mathbb{A}$, (EL).

If we now integrate (18) and the first of (16) by the different allowed schemes (section 3), we see that when $\mathbb{G}=0$ we truncate all linear terms, when $\mathbb{G}=\mathbb{A}_1$ we truncate the coupling terms, and when $\mathbb{G}=\mathbb{A}$ we deal with the exact linear solution.

We have seen from the last section that one-level implicit schemes are always stable and do not add parasitic modes. They have the disadvantage, however, of requiring an iteration process for calculation with an unspecified convergence rate. We have consequently added another scheme to our set of six (three implicit and three explicit) that is in effect an implicit scheme that can be solved without iteration. Consider the following two implicit techniques applied to the first of (16):

$$\psi_n^{t+\Delta t}-\psi_n^t=\tfrac{1}{2}\Delta t((\tilde{\boldsymbol{\Psi}}^*\mathbb{H}\boldsymbol{\Psi})^t+(\tilde{\boldsymbol{\Psi}}^*\mathbb{H}\boldsymbol{\Psi})^{t+\Delta t})$$

and

$$\psi_n^{t+\Delta t}-\psi_n^t=\tfrac{1}{4}\Delta t(\tilde{\boldsymbol{\Psi}}^{*t+\Delta t}+\tilde{\boldsymbol{\Psi}}^{*t})\mathbb{H}(\boldsymbol{\Psi}^{t+\Delta t}+\boldsymbol{\Psi}^t).$$

If we now combine these two schemes by taking the first twice and subtracting the second, we find

$$\psi_n^{t+\Delta t}-\psi_n^t=\frac{\Delta t}{2}(\tilde{\boldsymbol{\Psi}}^{*t}\mathbb{H}\boldsymbol{\Psi}^{t+\Delta t}+\tilde{\boldsymbol{\Psi}}^{*t+\Delta t}\mathbb{H}\boldsymbol{\Psi}^t). \tag{19}$$

On the assumption that the quantities ψ_n, $\boldsymbol{\Psi}$, $\boldsymbol{\Psi}^*$ are known at time t, equation (19) is linear in the terms at $t+\Delta t$ and may thus, in combination with a difference form of the type (19) used on (18), be solved for the variables at $t+\Delta t$. In terms of the vector $\mathbb{X}$ (five elements) defined as

$$\mathbb{X}\equiv\begin{pmatrix}\psi_n\\ \boldsymbol{\Psi}\\ \boldsymbol{\Psi}^*\end{pmatrix}, \tag{20}$$

the noniterative implicit scheme (IM) may be represented as

$$\mathbb{X}^{t+\Delta t}=\mathbb{R}_1^{-1}\mathbb{R}_2\mathbb{X}^t \tag{21}$$

where the matrices $\mathbb{R}_1$ and $\mathbb{R}_2$ are written

$$\mathbb{R}_1=\begin{bmatrix}1 & -\frac{\Delta t}{2}\tilde{\boldsymbol{\Psi}}^{*t}\mathbb{H} & \frac{\Delta t}{2}\tilde{\boldsymbol{\Psi}}^t\mathbb{H}\\ -\frac{\Delta t}{2}\mathbb{D}^t\boldsymbol{\Psi} & e^{-\mathbb{G}(\Delta t/2)}-\frac{\Delta t}{2}(\mathbb{A}-\mathbb{G}+\psi_n^t\mathbb{D}) & 0\\ -\frac{\Delta t}{2}\mathbb{D}^*\boldsymbol{\Psi}^{*t} & 0 & e^{-\mathbb{G}^*(\Delta t/2)}-\frac{\Delta t}{2}(\mathbb{A}^*-\mathbb{G}^*+\psi_n^t\mathbb{D}^*)\end{bmatrix}$$

and

$$\mathbb{R}_2=\begin{bmatrix}1 & 0 & 0\\ 0 & e^{\mathbb{G}(\Delta t/2)}+\frac{\Delta t}{2}(\mathbb{A}-\mathbb{G}) & 0\\ 0 & 0 & e^{\mathbb{G}^*(\Delta t/2)}+\frac{\Delta t}{2}(\mathbb{A}^*-\mathbb{G}^*)\end{bmatrix}.$$

Before proceeding to a discussion of the numerical calculations of the different schemes, let us consider the linearized coupled equations and the finite-difference solution to these equations. The linearization of equations (16) may be accomplished by assuming that $\psi_n \rightarrow \bar{\psi}_n = \text{constant}$ where it multiplies either ψ_α or ψ_β in the second equation of (16). The linearized equation may thus be expressed as

$$\dot{\mathbf{\Psi}} = \overline{\mathbf{G}}\mathbf{\Psi}$$
$$\overline{\mathbf{G}} \equiv \mathbf{A} + \bar{\psi}_n \mathbf{D} = i\begin{pmatrix} -\eta_\alpha & \bar{G}_{\alpha\beta} \\ \bar{G}_{\beta\alpha} & -\eta_\beta \end{pmatrix} \qquad (22)$$

where the elements of $\overline{\mathbf{G}}$ can be established from the definitions given in (16). The roots of $\overline{\mathbf{G}}$ are listed in table 2 together with the form of the modal matrix $\bar{\mathbf{S}}$ where

$$\overline{\mathbf{G}} = \bar{\mathbf{S}} i \bar{\mathbf{\Lambda}} \bar{\mathbf{S}}^{-1}$$

and $\bar{\mathbf{\Lambda}}$ is the root matrix of $\overline{\mathbf{G}}$. The solution to (22) may be written formally as

$$\mathbf{\Psi}(t) = e^{\overline{\mathbf{G}}t}\mathbf{\Psi}(t=0) = \bar{\mathbf{S}} e^{i\bar{\mathbf{\Lambda}}t} \bar{\mathbf{S}}^{-1}\mathbf{\Psi}(t=0). \qquad (23)$$

If the roots of $\overline{\mathbf{G}}$ are pure imaginary, no physical amplification will take place, and computational stability can be easily defined. The physical stability condition implied in $\overline{\mathbf{G}}$ has been discussed by Baer (1970) and need not be repeated here. We shall concern ourselves with physically stable situations.

Let us now apply the leapfrog scheme to (22) with the option that some of the linear terms may be extracted as we have done in (18). The appropriate form of (22), using the transformation (17), becomes

$$\dot{\chi} = e^{-\mathbf{G}t}(\overline{\mathbf{G}} - \mathbf{G})\mathbf{\Psi}$$

and (24)

$$\chi^{t+\Delta t} = \chi^{t-\Delta t} + 2\Delta t e^{-\mathbf{G}t}(\overline{\mathbf{G}} - \mathbf{G})\mathbf{\Psi}^t.$$

Here, **G** may take on any of the three values 0, $\mathbf{A}_1$, or **A**. Since we wish to compare the solutions of the second of (24) with (23), we return (24) to the variable $\mathbf{\Psi}$. Noting now that we may establish the roots of **G** and write a root matrix $\mathbf{\Lambda}$,

$$\mathbf{G} = \mathbf{S} i \mathbf{\Lambda} \mathbf{S}^{-1} \qquad (25)$$

where the roots and the modal matrix for different matrices **G** are listed in table 2, we find for (24) using (25) and (17)

$$\zeta^{t+\Delta t} = e^{2i\Lambda\Delta t}\zeta^{t-\Delta t} + 2\Delta t(\mathbf{S}^{-1}\overline{\mathbf{G}}\mathbf{S} - i\mathbf{\Lambda})\zeta^t$$
$$\zeta^t \equiv \mathbf{S}^{-1}\mathbf{\Psi}^t. \qquad (26)$$

Table 2.—*Values of the roots and modal matrices for var ious forms of G*

G	ζ	Λ_i	S
0	$\mathbf{\Psi}$	0	**I**
$\mathbf{A}_1$	$\mathbf{\Psi}$	ρ_α, ρ_β	**I**
A	$\mathbf{S}^{-1}\mathbf{\Psi}$	$-\frac{\rho_\alpha+\rho_\beta}{2} \pm \frac{1}{2}\{(\rho_\alpha-\rho_\beta)^2+4h_{\alpha\beta}h_{\beta\alpha}\}^{1/2}$	$-i\begin{pmatrix} \frac{\rho_\beta+\Lambda_1}{h_{\beta\alpha}} & \frac{\rho_\beta+\Lambda_2}{h_{\beta\alpha}} \end{pmatrix}$
$\overline{\mathbf{G}}$	-------	$\nu_{1,2} = -\frac{\eta_\alpha+\eta_\beta}{2} \pm \{(\eta_\alpha-\eta_\beta)^2+4\overline{\mathbf{G}}_{\alpha\beta}\overline{\mathbf{G}}_{\beta\alpha}\}^{1/2}$	$-i\begin{pmatrix} \frac{\eta_\beta+\nu_1}{\bar{G}_{\beta\alpha}} & \frac{\eta_\beta+\nu_2}{\bar{G}_{\beta\alpha}} \end{pmatrix}$

Since the elements of ζ are a linear combination of the elements of $\mathbf{\Psi}$, they will have the same solutions (roots). By the usual method of establishing an amplification matrix for multistep equations (Richtmyer 1957), we define the vector ξ as

$$\xi^{t+\Delta t} = \zeta^t,$$

and we get the solution to (26) in the form

$$\begin{pmatrix} \zeta \\ \xi \end{pmatrix}^{t+\Delta t} = \begin{pmatrix} 2\Delta t(\mathbf{S}^{-1}\overline{\mathbf{G}}\mathbf{S}) - i\mathbf{\Lambda} & e^{2i\Lambda\Delta t} \\ \mathbf{I} & 0 \end{pmatrix} \begin{pmatrix} \zeta \\ \xi \end{pmatrix}^t \qquad (27)$$

where **I** is the unit matrix. The root equation for the amplification matrix in (27) is given as

$$\begin{vmatrix} 2i\Delta t(\mathbf{S}^{-1}\bar{\mathbf{S}}\bar{\mathbf{\Lambda}}\bar{\mathbf{S}}^{-1}\mathbf{S} - \mathbf{\Lambda}) - \lambda\mathbf{I} & e^{2i\Lambda\Delta t} \\ \mathbf{I} & -\lambda\mathbf{I} \end{vmatrix} = 0 \qquad (28)$$

that is, in general, a fourth-order equation in the roots. Two of these roots are physically real; the other two are parasitic. The real roots should be compared with the roots of the exact solution, $e^{i\bar{\Lambda}\Delta t}$. So long as Δt remains within the limits of computational stability, the roots will have amplitude of unity, and we may therefore consider only the phase angles. Thus if the roots of (28) are

$$\lambda_j = e^{i\theta_{\lambda j}}, \qquad (29)$$

we may compare the phase angles for the finite-difference solution to those of the exact solution by the ratio $\theta_{\lambda j}/\nu_j\Delta t$. These ratios are shown for each of the approximations $\mathbf{G}=0$, $\mathbf{A}_1$, and **A** on figure 4 plotted against the nondimensional time unit Δt where time has been nondimensionalized by the earth's rotation rate. The data used in determining the roots was taken from case CA and is given in table 3. The values of the frequencies $\nu_{1,2}$ are $\nu_1=0.310$ and $\nu_2=0.042$. We shall have occasion to compare these values to the exact frequencies of the nonlinear solution in the next section.

The phase errors for the approximation $\mathbf{G}=0$ may be readily determined since (28) reduces to the equation

$$\lambda^2 - 2i\Delta t\nu_j\lambda - 1 = 0$$

from which we see that the phase angles $\theta_{\lambda j}$ (equation 29) are given by

$$\theta_{\lambda j} = \sin^{-1}(\nu_j\Delta t),$$

a result identical to the one arrived at in section 3 for uncoupled systems. The stability criterion and phase error at the stability point are

$$\Delta t = \frac{1}{\nu_j}, \left(\frac{\theta_{\lambda j}}{\nu_j\Delta t}\right)_{critical} = \pi/2.$$

Returning to the discussion of figure 4, we see that removal of the uncoupled terms ($\mathbf{G}=\mathbf{A}_1$) leads to a much more stable calculation with considerably lower errors in

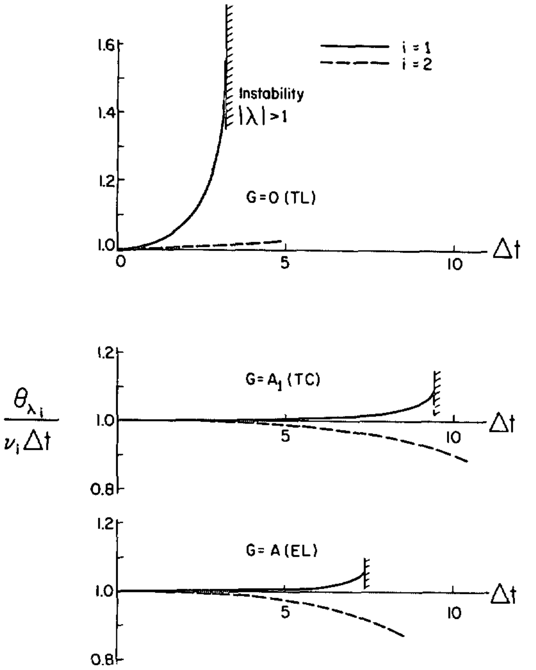

FIGURE 4.—Phase errors for the coupled system using the leapfrog scheme and the three linear truncation methods TL, TC, and EL.

TABLE 3.—*Numerical values of the variables in three different sets of conditions used in this study*

Constants in eq. (15)		*Case* CA	*Case* CB	*Case* CC
	a_n	−0.09788005	−0.69019528	−4.8903939
	ρ_α	−.26691770	.21502294	0.79837156
	ρ_β	−.03839707	.86136911	.38763314
	$h_{\alpha\beta}$	−.09899117	.27432320	1.1105865
	$h_{\beta\alpha}$	−.07127364	.23888073	0.59288305
	$g_{\alpha\alpha}$	−.08220212	.19160456	−1.0310522
	$g_{\alpha\beta}$	−.03050807	.43151307	7.2731965
	$g_{\beta\alpha}$	−.12767626	−.41340017	4.1460202
	$g_{\beta\beta}$	−.01396861	.34083072	3.2491569
Initial values	ψ_n	−0.60497847	0.80465985	−0.28630513
	ψ_α	.63421748	.42852365	.10114551
	ψ_β	−.25891820	.10713091	.07680246
Solutions of eq. (15)				
Energy variations	zonal	0.374→0.244	0.700→0.271	0.200→0.292
(normalized)	α-wave	.402→.613	.245→0.627	.464→.076
	β-wave	.223→.143	.055→.102	.336→.632
Energy exchange period (days)		3.452	1.470	0.508
Wave periods observed in exact solutions (days)		3.24 52.	1.29 10.3	0.527 —
Wave frequencies from linearized equations: $\nu_{1,2}$		0.3097 .0421	−0.6407 −.1040	−1.903 0.0404
Corresponding wave periods (days)		3.23 23.8	1.56 9.6	.526 24.75

phase. Total truncation (G=0) is clearly the worst case, whereas including an exact treatment of the coupling terms in A_2 does not improve the stability or phase errors; in fact, the extraction of more exact information in this case creates larger truncation errors. It should be noted that these results refer to a particular set of initial conditions and are subject to change for different conditions. However, we may conclude with some confidence that the exact treatment of uncoupled linear terms will yield solutions to the nonlinear equations with less error for a given truncation element Δt. We shall not consider the linearized equations for the other truncation schemes but proceed directly to the numerical calculation of the nonlinear equations.

5. NUMERICAL CALCULATIONS

Three different sets of initial conditions were used to test the truncation schemes; they are listed in table 3 and are denoted respectively as cases CA, CB, and CC. Since the exact solutions to (16) are known, any variable determined from a numerical integration will be represented normalized by its exact value. For each case of initial data, the three methods (TL, TC, and EL) for dealing with the linear terms were applied, and three different time increments (Δt) were used; the scales of the time increments were determined from the characteristic frequencies of the cases. All seven truncation schemes discussed in section 3 as having satisfactory linear properties were tested and are listed in table 4.

Let us now concentrate our attentions on the features of case CA that was integrated numerically in excess of 51 days. Since this case has an exact nonlinear exchange period of 3.452 days, the integration period should be long enough to highlight important errors. The exact frequencies—and there are two because two wave components, ψ_α, ψ_β exist—are $\nu_1=0.309$, $\nu_2=0.0192$. The first of the two frequencies calculated by linear theory (section 4) compares remarkably well with the first exact frequency, but the second is more than twice as large. However, because of the difference in magnitude of these frequencies, the first (larger) frequency will essentially determine the stability criterion. Using the first frequency and the linear (uncoupled) solutions for the different schemes developed in section 3, we list in table 4 the stability condition Δt_{max}, which the linear theory would indicate.

A common procedure for establishing stability and truncation errors is to investigate the development in time of some integral property of the system—generally conservative—as was done by both Lilly (1965) and Young (1968). For simple atmospheric flow problems, energy is the logical choice, although Young also included the vorticity. To indicate the behavior of the total energy of our solution (case CA) with time, we have prepared table 4 in which we describe the total energy (conserved in the exact solution) for the different truncation schemes, different truncation intervals Δt=2.07, 4.14, or 8.28 hr, and different treatment of the linear terms. The energies have been listed after 51.77 days unless an oscillation occurs, in which case its range is tabulated. As indicated above, we have also listed the stability condition based on linear theory.

Unquestionably, the stability properties are well described by the total energy and correspond to those anticipated from linear theory. Where damping is pre-

Table 4.—*Total energy normalized by the exact value for case* CA *after 51.77 days for seven schemes, three different time steps, and different treatment of linear terms together with the linear stability criterion. In case of oscillation, the range is tabulated.*

Scheme stability Δt=(hours)→	TL 2.07	 4.14	 8.28	TC 2.07	 4.14	 8.28	EL 2.07	 4.14	 8.28
E_{11}	0.999	0.991	0.934	----------	0.998	0.988	0.999	0.968	0.889
				1.000					
$\Delta t \leq 12$ hr	1.001	1.006	1.098	----------	1.001	1.006	1.003	1.045	81.881
E_{03}									
$\Delta t \leq 5$ hr	0.993	0.822	Overflow	1.000	0.989	0.855	0.999	0.979	0.801
			(11 days)		.991	.857		.980	.812
E_{33}									
$\Delta t \leq 5$ hr	1.000	0.999	Overflow	0.998	0.988	Overflow	0.878	Overflow	
		1.002	(8 days)	1.002	1.013	(32 days)	1.409	(50 days)	
IM									
$\Delta t \leq \infty$	----------	----------	----------	1.000	1.000	1.000	0.999	0.995	0.982
					1.002	1.007	1.000	1.000	1.000
I_{01}	1.000	1.000	0.999	1.000	1.000	0.999	0.999	0.998	0.999
$\Delta t \leq \infty$		1.001	1.005		1.001	1.003	1.000	1.001	1.011
I_{13}	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	0.999
$\Delta t \leq 21$ hr			1.001						1.000
I_{35}	1.000	1.000	0.999	1.000	1.000	0.999	1.000	1.000	0.999
$\Delta t \leq 16$ hr			1.003			1.000			1.001

dicted, as in scheme E_{03}, the tabular values are in agreement. Where parasitic oscillations are anticipated (E_{11}), they appear in the table. Further expected results show that the solutions deteriorate for increased Δt and that implicit schemes are generally superior (for given Δt) than explicit ones. A further observation, not previously investigated, is the improvement of the solution from TL to TC; that is, when the uncoupled linear terms are treated exactly. If, however, one proceeds to treat all linear terms exactly (EL), the results appear somewhat less stable, as expected from the linear analysis (section 4).

The above information is indeed valuable; however, it must be emphasized that *the behavior of the total energy with time is not necessarily an indicator of the behavior of the detailed character of the solution.* As we shall see, the individual amplitudes of the wave components may be seriously in error with no indication from the total energy. Moreover, the phase angles and wave velocities of the components from the truncated calculations may have no relation to the true solution, although the total energy is well conserved. To establish this fact, among others, we shall proceed to a detailed discussion of the calculations.

The component amplitudes that make up the total energy in our equations may be represented when we describe the truncated value normalized by the exact value from equations (15) as

$$\begin{aligned}
&A \text{ energy} \equiv 2c_\alpha\psi_\alpha\psi_\alpha^* \text{ (truncated)}/2c_\alpha\psi_\alpha\psi_\alpha^* \text{ (exact)},\\
&B \text{ energy} \equiv 2c_\beta\psi_\beta\psi_\beta^* \text{ (truncated)}/2c_\beta\psi_\beta\psi_\beta^* \text{ (exact)},\\
&Z \text{ energy} \equiv \Sigma c_\gamma\psi_\gamma^2 \text{ (truncated)}/\Sigma c_\gamma\psi_\gamma^2 \text{ (exact)},\\
\text{and} \qquad &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (30)\\
&T \text{ energy} \equiv \text{total energy}.
\end{aligned}$$

The time variation of the three energy components presented in (30) have been plotted for Δt=4.14 hr for all seven schemes listed in table 4 for both the TL and TC conditions based on data from case CA in figure 5. We have selected to discuss the TL condition because it is by far in most common usage and the TC condition for comparison. From a superficial view of figure 5, one is immediately impressed with the sizable errors in some of the schemes, a fact not established from table 3. These errors have a regular period that is given by the first (largest) frequency, ν_1=0.309. One must conclude, therefore, that the energy components cancel their errors on summation. A further observation is the remarkable improvement in the calculations (reduction in error) by use of the TC condition. Although this condition has been in computational use with higher order systems for some time (Baer 1964), its virtues had not been investigated in any detail.

Of the explicit schemes tested, E_{33} is by far the best with regard to truncation, showing almost no errors during the entire integration period for $\Delta t \simeq 4$ hr. However, in terms of its utility as a computation scheme, we must refer back to table 4 that elucidates its limited stability region ($\Delta t \leq 5$ hr). Scheme E_{03} shows errors in excess of 50 percent in the energy components and describes the anticipated damping with time, but only in the α wave. The leapfrog scheme also shows large error excursions, but they are cut back dramatically by the TC condition.

Although table 4 indicates no significant errors for the implicit schemes, figure 5 clearly does not corroborate this interpretation. Scheme I_{01} has errors as large as 50 percent in the components for the TL condition; they are, however, almost completely eliminated when the TC method is applied. An unfortunate and unexpected result of the tests is the poor quality of the IM computation. While the TL results are not available, the TC results suggest that this scheme is inferior to the others described on figure 5 (another observation not anticipated from the total energy information of table 4). Schemes I_{13} and I_{35}

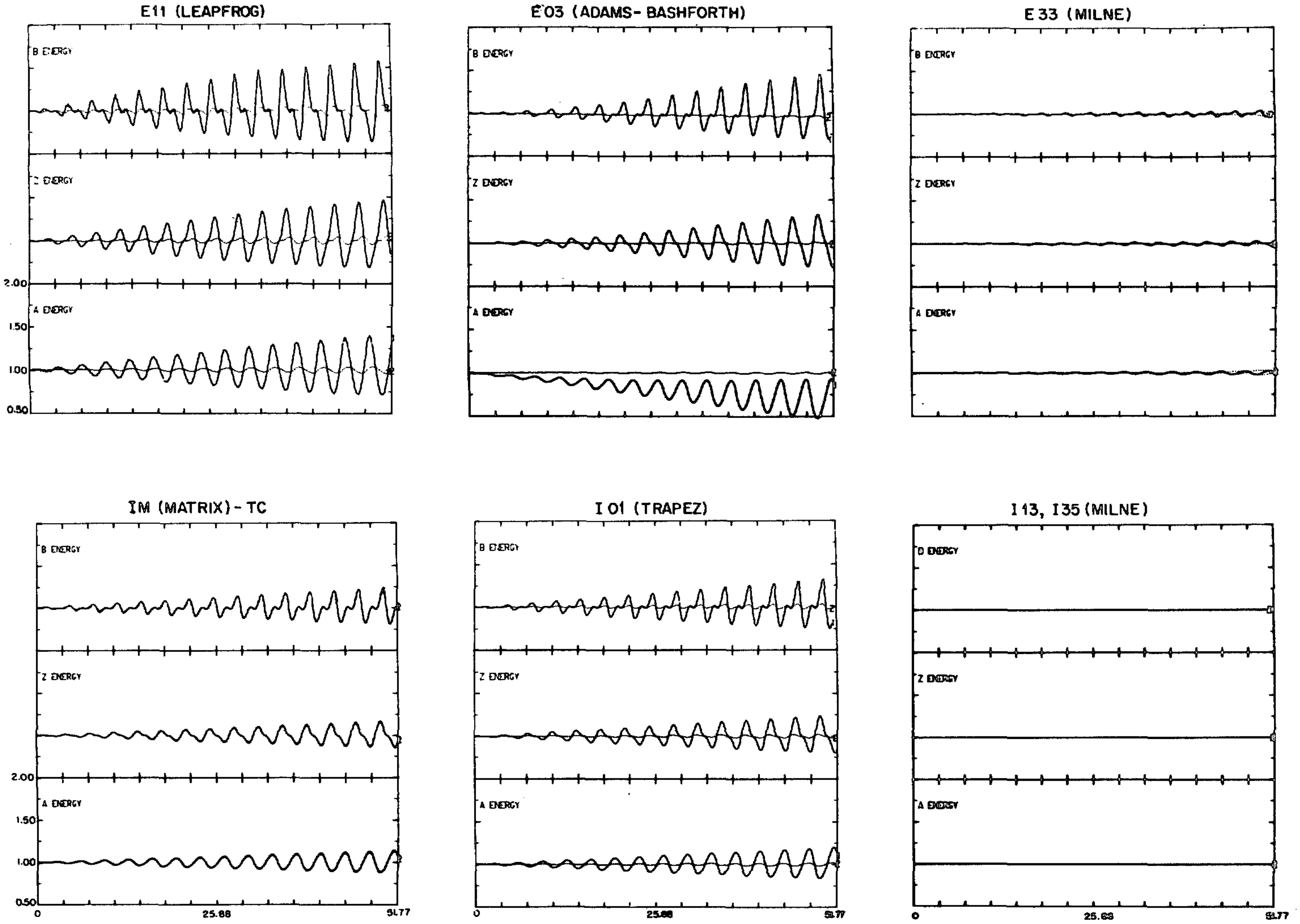

FIGURE 5.—Energy in the zonal and α and β waves as a function of time in days (abscissa) normalized by their exact values for the seven schemes that had favorable linear properties. Solid curves represent TL condition and dotted curves are for TC, both for Δt=4 hr.

have been plotted on the same chart since neither has any measurable error in the energy components over the total integration period for $\Delta t \simeq 4$ hr. They are clearly superior schemes, but I_{13} should be preferred, both because of its better stability condition (table 4) and its ease of computation.

As indicated above, a striking feature described by figure 5 is the improved computation for the TC condition. Because this involves the exact treatment of part of the linear contribution, one might anticipate that the exact treatment of all the linear terms (EL) might further improve the calculated results. That this reasoning is incorrect has already been suggested by the deterioration of the stability criterion for the leapfrog scheme using the EL condition, seen from the total energy in table 4. Since E_{11} shows this feature most strongly of all the schemes, we describe on figure 6 the different energy components with time for E_{11}, Δt=4.14 hr using both the TC and EL conditions; the comparison of TL to TC is evident from figure 5. None of the schemes show improved computation using the EL method, but most give results comparable to the TC calculation. Most remarkable is the instability that is set up in the E_{11} scheme using the EL method, a result not anticipated from the linear analysis of section 4 (fig. 4), wherein the stability condition for the EL calculation was superior to the TL method. We find here, therefore, a purely nonlinear phenomenon, not predictable by linearization. However, this observation is not systematic with regard to all the schemes and does *not* appear for E_{03}.

The error in the energy components as a function of Δt is described by figure 7. Here we show both E_{11} and I_{01} using the TL method for the three times, Δt=2.07, 4.14, and 8.28 hr. We have chosen E_{11} and I_{01} because they are the most frequently used schemes in the explicit and implicit groups, respectively. Nevertheless, all schemes tend to show a similar deterioration of the result with increased Δt, although the higher level implicit schemes

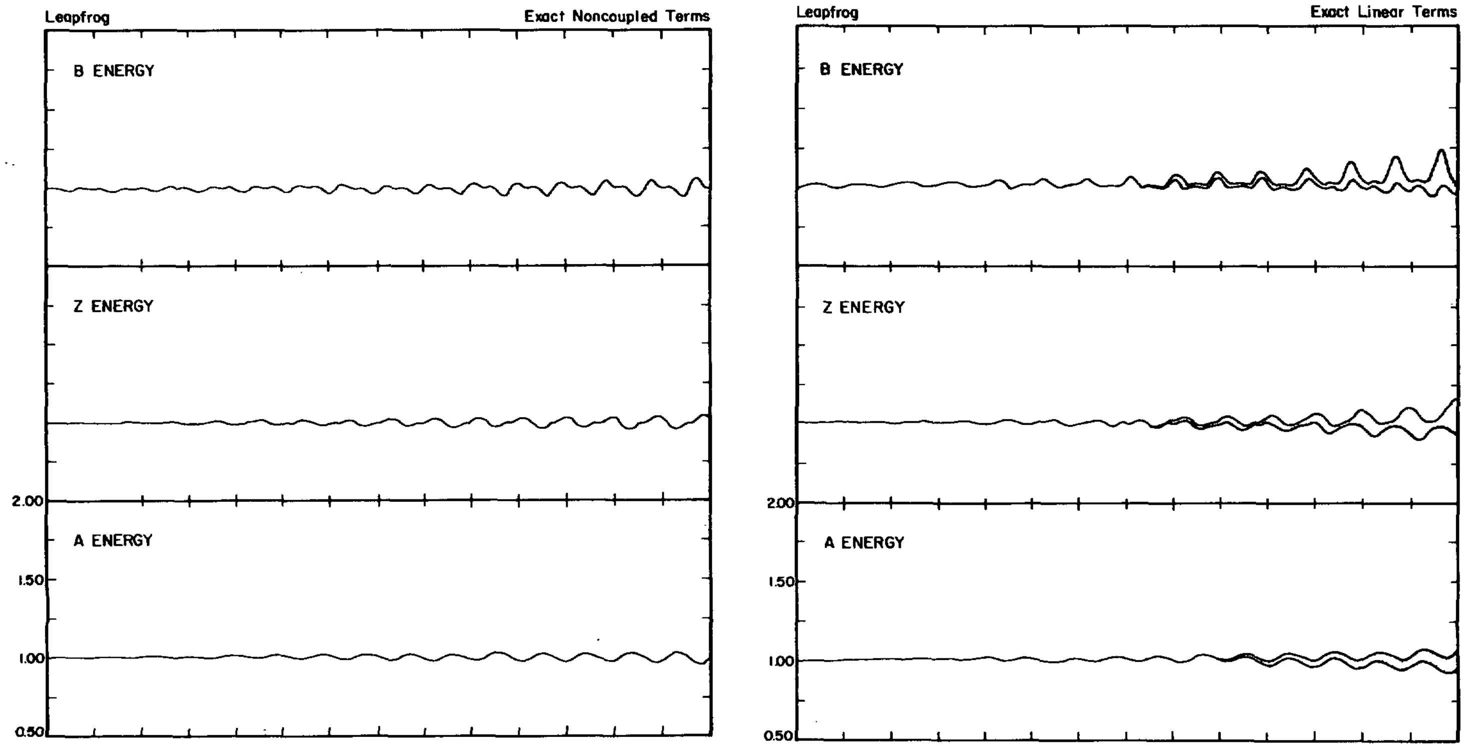

Figure 6.—Component energies calculated using the leapfrog scheme with $\Delta t \simeq 4$ hr, showing the difference between the TC (exact noncoupled) and EL (exact linear) methods.

(I_{13} and I_{35}) have extremely small errors for $\Delta t \simeq 8$ hr. The failure of the total energy to indicate the errors in the components is plainly evident from this figure. An interesting feature of the leapfrog scheme that is apparent for $\Delta t \simeq 8$ is the larger error period, a modulation effect caused by the parasitic mode; this phenomenon has been observed and discussed in the past (Baer 1961). An indication of the component errors seen on figure 7 may be available from linear theory through the phase errors. When $\Delta t \simeq 2$, the phase errors (fig. 3) are almost indetectable; whereas when $\Delta t \simeq 8$ ($\nu\Delta t \simeq 2/3$), both the leapfrog and trapezoidal schemes show sizable phase errors. It is interesting to note that for the latter truncation the Milne scheme (I_{13}) has almost no linear phase error and correspondingly no nonlinear computational errors.

Despite the appearance of large errors in the energy components, there exist periodic times at which the computed solutions describe the exact solution with great accuracy. One might thus be led to the conclusion that the numerical integrations will give satisfactory results at selected times (periodic) for all time, to be determined by the highest characteristic mode of oscillation (available from linear theory). Such reasoning, in analogy with the conclusions drawn from the behavior of the total energy only, is based on incomplete information and is unfortunately incorrect. The missing information are the phase angles of the α and β waves, both of which are time dependent; their time dependence may be described by the real part of the stream components ψ_α, ψ_β and we present them as

$$A\text{-wave} \equiv \mathrm{Re}\psi_\alpha\ (t)$$

and (31)

$$B\text{-wave} \equiv \mathrm{Re}\ \psi_\beta\ (t).$$

In figure 8, we show the phase properties for the E_{11} and I_{01} schemes for the time steps $\Delta t = 4.14$, 8.28 hr, using the TL method. By comparing the computed values of the two wave components as given in (31) to the exact values, we see that after 50 days the A-wave is significantly out of phase with the exact value. For $\Delta t \simeq 8$ hr, the phase error is almost 180° in both schemes, whereas the error in the B-wave is negligible. This error grows with time, and the consequent solution therefore becomes less and less reliable. Having now established an almost insurmountable obstacle to these numerical integration schemes (the multistep implicit schemes I_{13}, I_{35} do not exhibit discernible phase errors for the time steps utilized), we observe that no apparent phase errors occur if we use the TC or EL condition. The interpretation of this correction must be based on the fact that the uncoupled linear terms include most of the high-frequency phase properties and therefore cannot be successfully truncated. Although many of the schemes exhibit the phase characteristics outlined above, the implicit matrix scheme (IM) is nonconformist. With $\Delta t = 8.28$ hr, figure 8 shows the phase properties for both

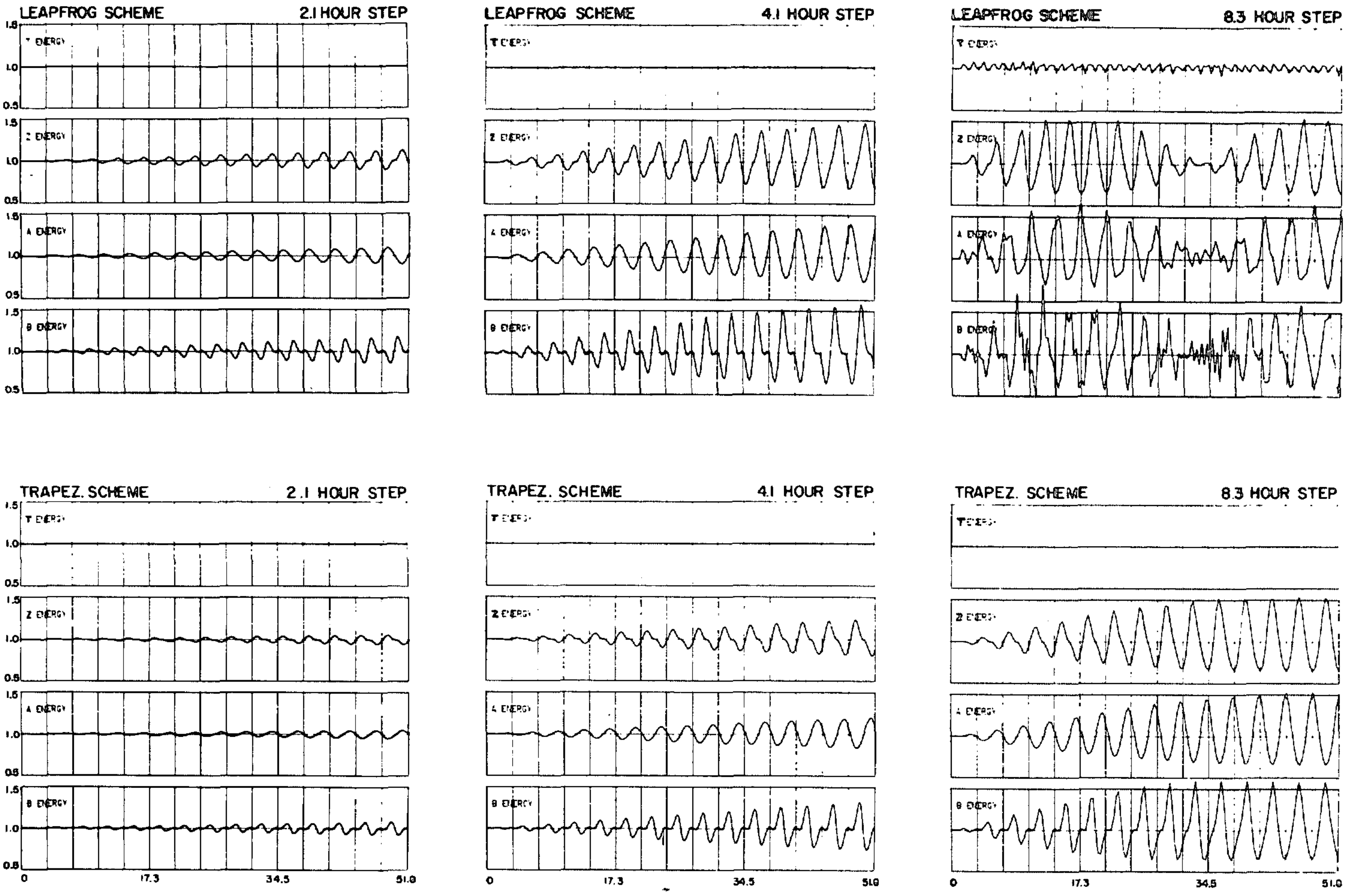

FIGURE 7.—Component energies calculated using the E_{11} and I_{01} schemes with the TL condition, showing the effect of increasing Δt.

the TC and EL methods of the IM scheme and highlights the phase errors, here primarily in the long period of the *B*-wave.

Figures 9 and 10 (to lend some credence to generalizations from the above observations based only on case CA) describe the behavior in time of the energy components for data from cases CB and CC, respectively (numerical values to be found in table 3). The results described are based on the TL method, and the time increments have been selected on the basis of the characteristic frequencies (table 3). All the features that these figures can describe are similar to those discussed for case CA. Errors increase for increased Δt, total energy is well conserved whereas the component errors are large, a modulation period appears in the leapfrog scheme, and the Milne (I_{13}) scheme is extremely accurate. We have found that the other features discussed in detail for case CA shows similar properties for cases CB and CC, and we shall consequently not reproduce these results here; we shall assert, however, based on figures 9 and 10, that the computational properties of the different schemes tested and discussed in this section are applicable to a wide variety of initial conditions.

6. CONCLUSION

The solution of the nonlinear equations that describe atmospheric flow (among others) by numerical means is today a commonplace event. These equations (given a set of initial values) are frequently integrated in time for long periods. It is therefore imperative that an integration scheme be chosen that is not only stable but also has negligible truncation errors so that the true solution is not obscured. The development of the "spectral" approach allows this solution to be carried out in time alone, thereby bypassing the space truncation influence. Moreover, the reduction of the spectral equations to low-order form, with their known solutions, enables us to test directly the validity and accuracy of any truncation technique.

Since a wide variety of schemes exist and have been applied, it is desirable to find a general method whereby such schemes may be systematically presented for testing. We have developed such a method based on finite-difference polynomial interpolation and have shown that many of the more common schemes—both implicit and explicit—are incorporated in our presentation. A number of the lower level schemes have been tested on a simple linear

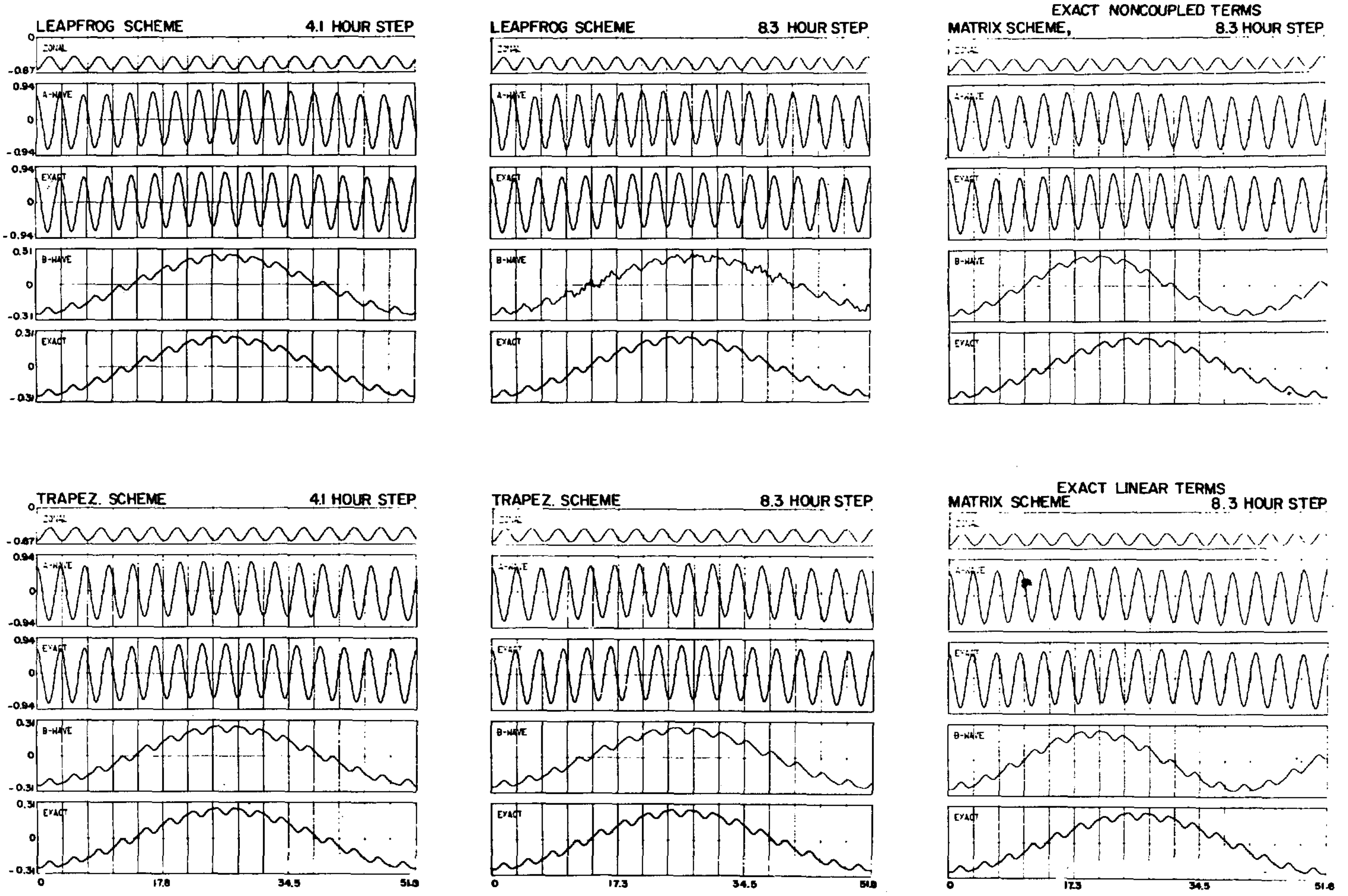

FIGURE 8.—Computed and exact values of the wave components Re ψ_α, Re ψ_β for the leapfrog, trapezoidal, and matrix (IM) schemes, for various Δt and linear conditions, showing development of phase errors.

wave equation, and those with the most favorable qualities (best stability condition and least truncation) have been selected for testing with a low-order nonlinear spectral system. Included in this group is an implicit method that is not a member of the general set, but is interesting because it does not require iteration.

The low-order system is of particular interest as it involves both linear (coupled and uncoupled) and nonlinear effects. Linear terms may be handled without truncation, and a procedure whereby these terms are removed from the equations may have some impact on the numerical solution of the remaining purely nonlinear equations. An indication that the truncation errors are modified by such elimination is suggested from the solution of the linearized low-order equations, both exactly and with finite-difference methods.

The comparison of the truncated solutions to the exact ones yields some interesting observations. Whereas it has been common to estimate truncation errors of an integration from the behavior of conservative integral properties, our results indicate that only stability can be discussed in this way. The amplitudes of functional variables in our nonlinear system showed wild deviations (errors) at times during the numerical integration, but the conservative property (energy) was well conserved; this was caused by a cancellation of the individual amplitude errors. One must conclude that the conservation of integral constraints in a numerical calculation is not sufficient to justify confidence in the results. Furthermore, the satisfactory prediction of amplitudes is also not sufficient; one must also assure the accurate calculation of the phase angles.

Linear theory seems to yield satisfactory information about the computational stability of our nonlinear system, as may be seen from the development of the conservative property; and the linear phase errors (for any scheme) are indications of errors in the amplitudes of the dependent variables. Nonlinear phase errors, which are pronounced for the explicit schemes, may be removed by the exact consideration of the uncoupled linear terms of the nonlinear equations; the latter technique also reduces the amplitude errors significantly. As might have been anticipated, reduction of the truncation interval, Δt, will yield improved solutions.

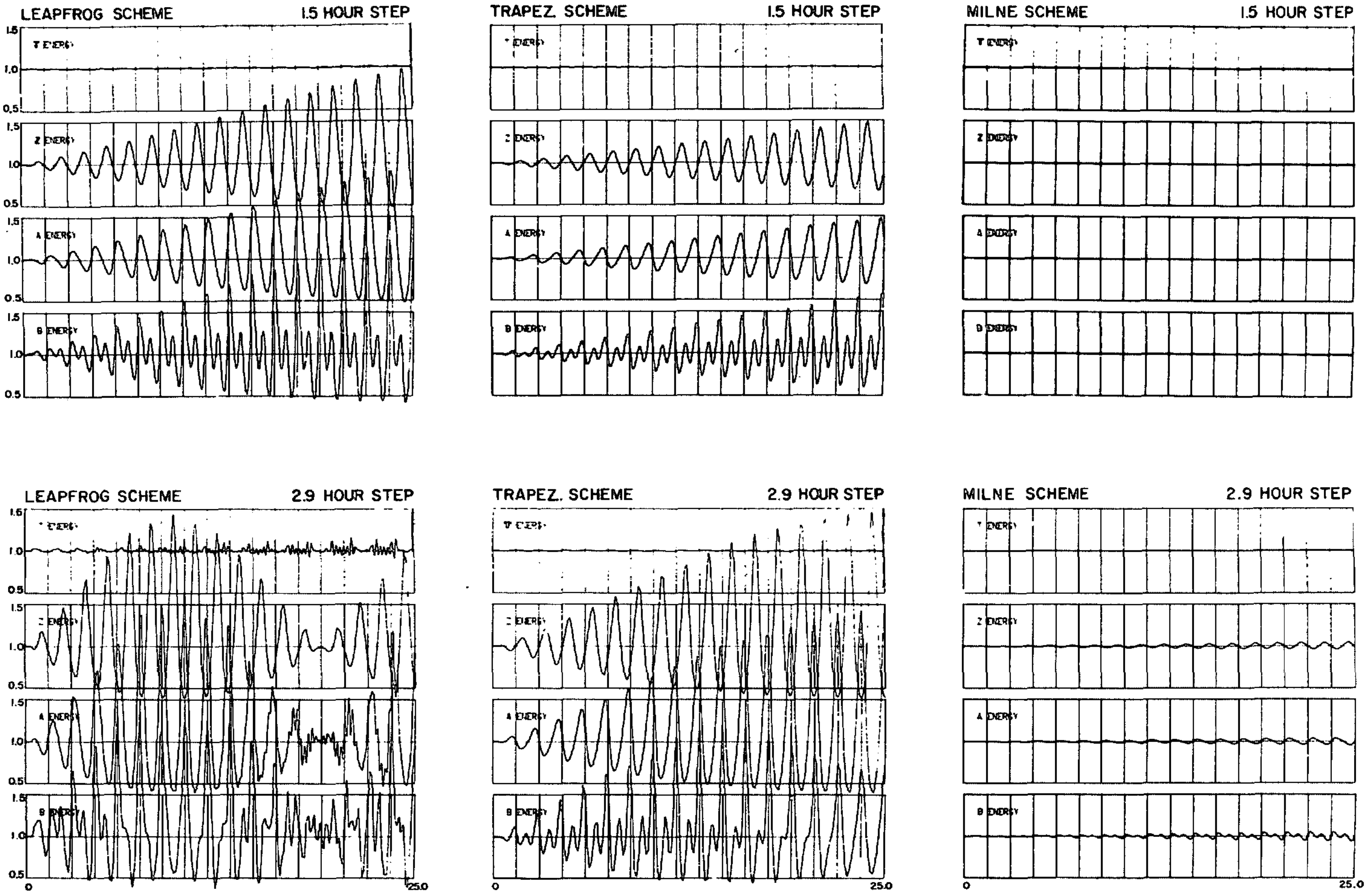

FIGURE 9.—Component energies calculated using the leapfrog, trapezoidal, and Milne (I_{13}) schemes using the TL condition for the two time steps $\Delta t=1.5$, 2.9 hr for case CB (see table 3).

As a consequence of our calculations, it would be most advisable to select a truncation increment (Δt) substantially less than the critical one determined from linear analysis, if truncation errors are to be minimized. Moreover, to avoid phase errors, one must remove any uncoupled linear terms from the equations by a linear transformation involving the exact solution of such terms. Finally, if computation time is not a serious consideration, an implicit method should be selected in preference to an explicit one. Multistep methods, although they involve more parasitic solutions, seem to yield superior results. If, for reasons of economy and speed, an explicit scheme is chosen, a technique denoted as "restart" that begins a new calculation periodically from the mean data at the restart time appears to reduce high-frequency amplifying parasitic oscillations, but other truncation properties of this procedure have not been evaluated.

ACKNOWLEDGMENTS

Computer time for this study was made available by the computer facility of the National Center for Atmospheric Research and by the computer center at Colorado State University. This research has been supported by the Atmospheric Sciences section of the National Science Foundation under Grant No. GA-761 to Colorado State University.

REFERENCES

Baer, Ferdinand, "The Spectral Vorticity Equation," Ph. D. thesis, Department of Geophysical Sciences, The University of Chicago, **1961**, 97 pp.

Baer, Ferdinand, "Integration With the Spectral Vorticity Equation," *Journal of the Atmospheric Sciences*, Vol. 21, No. 3, May **1964**, pp. 260–276.

Baer, Ferdinand, "Analytic Solutions to Low-Order Spectral Systems," *Archives for Meteorology, Geophysics, and Bioclimatology Serie A, Meteorology and Geophysics*, Vol. 19, No. 3, Springer-Verlag, Vienna, **1970**.

Henrici, P., *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley and Sons, Inc., New York, **1962**, 407 pp.

Henrici, P., *Elements of Numerical Analysis*, John Wiley and Sons, Inc., New York, **1964**, 328 pp.

Hildebrand, F. B., *Introduction to Numerical Analysis*, McGraw-Hill Book Co., Inc., New York, **1956**, 510 pp.

Kurihara, Yoshio, "On the Use of Implicit and Iterative Methods for the Time Integration of the Wave Equation," *Monthly Weather Review*, Vol. 93, No. 1, Jan. **1965**, pp. 33–46.

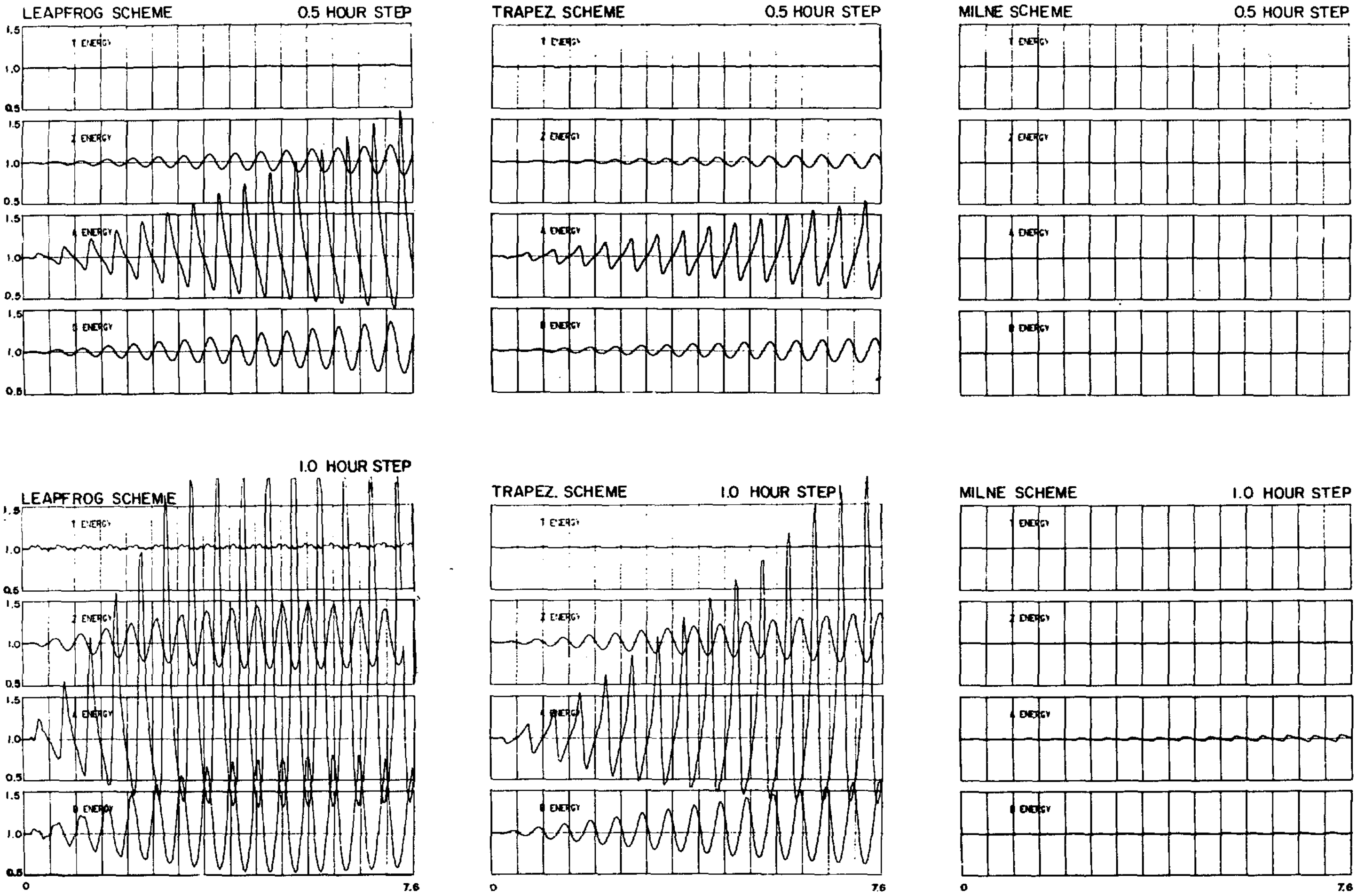

FIGURE 10.—Component energies calculated using the leapfrog, trapezoidal, and Milne (I_{13}) schemes using the TL condition for the two time steps $\Delta t = 0.5$, 1.0 hr for case CC (see table 3).

Lilly, Douglas K., "On the Computational Stability of Numerical Solutions of Time-Dependent Non-Linear Geophysical Fluid Dynamics Problems," *Monthly Weather Review*, Vol. 93, No. 1, Jan. **1965**, pp. 11–26.

Lorenz, Edward N., "Maximum Simplification of the Dynamic Equations," *Tellus*, Vol. 12, No. 3, Aug. **1960**, pp. 243–254.

Milne, W. E., *Numerical Calculus*, Princeton University Press, Princeton, N.J., **1949**, 393 pp.

Platzman, George W., "The Spectral Form of the Vorticity Equation," *Journal of Meteorology*, Vol. 17, No. 6, Dec. **1960**, pp. 635–644.

Platzman, George W., "The Analytical Dynamics of the Spectral Vorticity Equation," *Journal of the Atmospheric Sciences*, Vol. 19, No. 4, July **1962**, pp. 313–328.

Richtmyer, R. D., *Difference Methods for Initial Value Problems*, Interscience Publishers, Inc., New York, **1957**, 238 pp.

Silberman, Isadore, "Planetary Waves in the Atmosphere," *Journal of Meteorology*, Vol. 11, No. 1, Feb. **1954**, pp. 27–34.

Young, John A., "Comparative Properties of Some Time Differencing Schemes for Linear and Nonlinear Oscillations," *Monthly Weather Review*, Vol. 96, No. 6, June **1968**, pp. 357–364.

[*Received December 17, 1969; revised February 24, 1970*]